When Policies Cannot Be Retrained: A Unified Closed-Form View of Post-Training Steering in Offline Reinforcement Learning

Elias Hossain, Mohammad Jahid Ibna Basher, Ivan Garibay, Ozlem Garibay, Niloofar Yousefi

Complex Adaptive Systems Laboratory, University of Central Florida



2026

project: eliashossain001.github.io/poe-steering

code: github.com/eliashossain001/poe_composition

The objective moved. The actor cannot.

An offline RL actor has been trained and approved. Then the deployment objective is revised, and changing the actor is not an option: data access is limited, retraining is expensive, or every policy change needs governance and revalidation.

No new interaction

The environment cannot be queried at deployment.

No new data

Only the original offline dataset is available.

No retraining

The actor’s parameters stay frozen.

Goal: steer the frozen actor toward the new objective while staying conservative relative to the validated policy.

Method: steer by composing with a goal prior

$$\pi'(a \mid s, g) \propto \pi(a \mid s)^{\alpha} \cdot \rho(a \mid s, g)^{1-\alpha}$$

Two experts, one knob

π · frozen diagonal-Gaussian actor, trained once by behavioral cloning; **π′** is the deployed policy

ρ · goal-conditioned prior, trained on the same offline data; it never queries a critic

α · $\alpha \to 1$ returns the frozen actor; $\alpha \to 0$ returns the prior on the actor's support

Gaussian closed form

$$\Sigma'^{-1} = \alpha\, \Sigma\pi^{-1} + (1-\alpha)\, \Sigma\rho^{-1}$$

$$\mu' = \Sigma' \left[\, \alpha\, \Sigma\pi^{-1}\, \mu\pi + (1-\alpha)\, \Sigma\rho^{-1}\, \mu\rho \,\right]$$

Precision-weighted: a confident prior moves the action, a diffuse one barely does.

Finding 1: PoE(α) and KL-Reg(β) are the same policy

$\beta = \alpha / (1 - \alpha)$

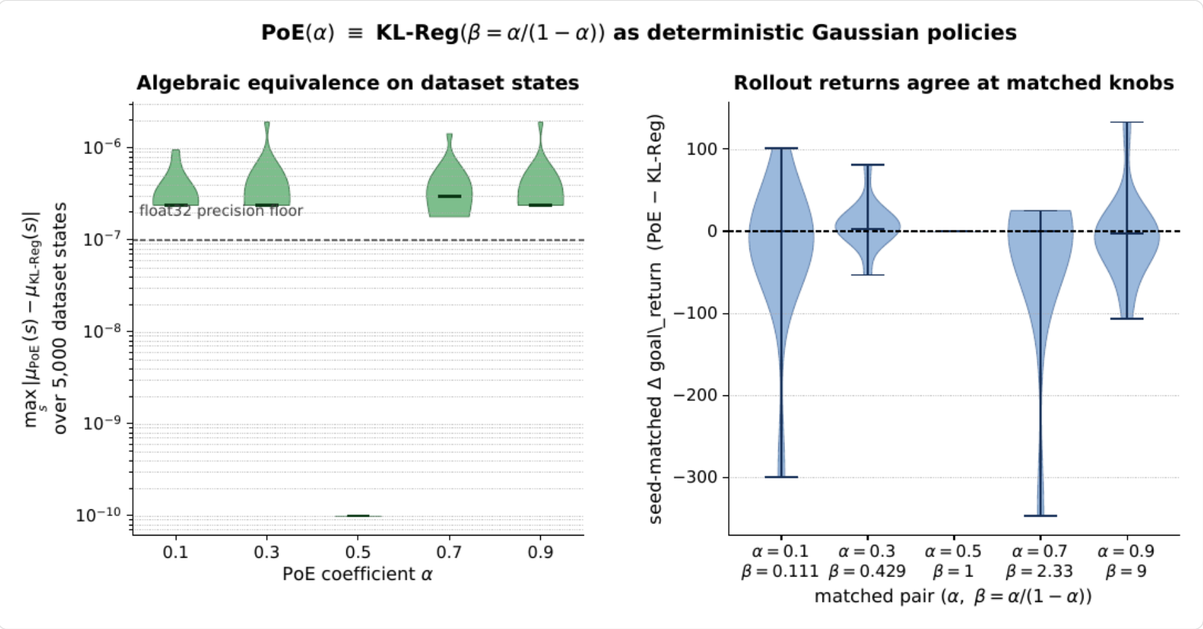
KL-regularized adaptation, argmin E[−log ρ] + β KL(π′ ‖ π), gives π* ∝ π · ρ^(1/β). Its natural parameters are a scalar multiple of PoE's, so the means coincide and the covariances differ by 1/α, which cancels under mean-action deployment.

≤ 2×10⁻⁶

max mean gap, 5,000 states

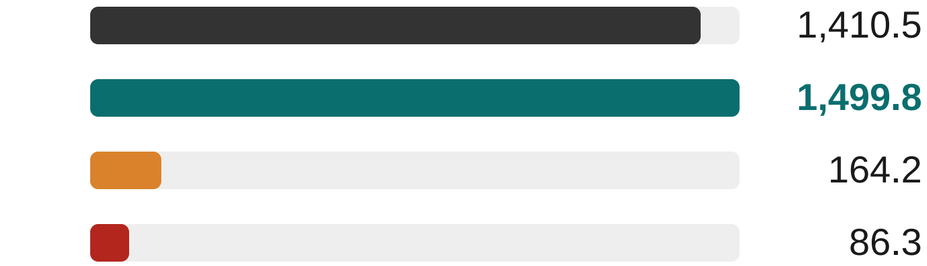
12 / 12

bit-identical rollouts at α = 0.5

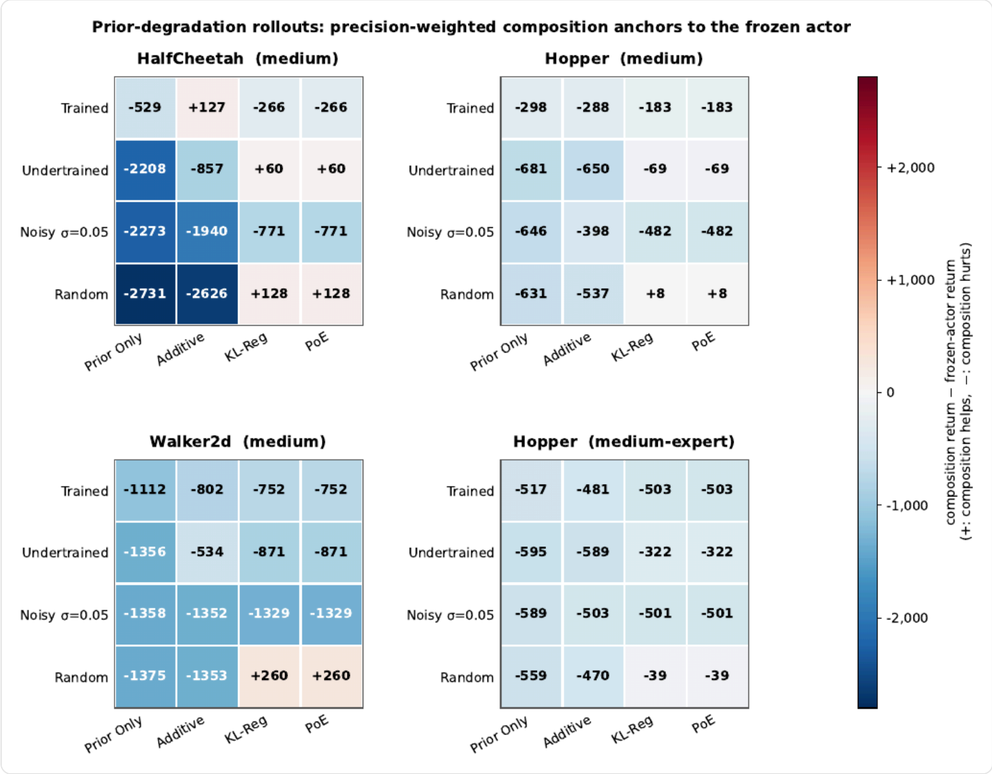


Finding 2: with a bad prior, PoE stays with the actor

Mean goal-weighted return, random prior



The order holds under noisy and undertrained priors. A degraded prior has low precision, so the PoE mean shrinks back toward the actor; additive mixing carries the damaged mean straight through.



Finding 3: with a good prior, help, tie, or hurt

4
Help

4
Tie (Frozen)

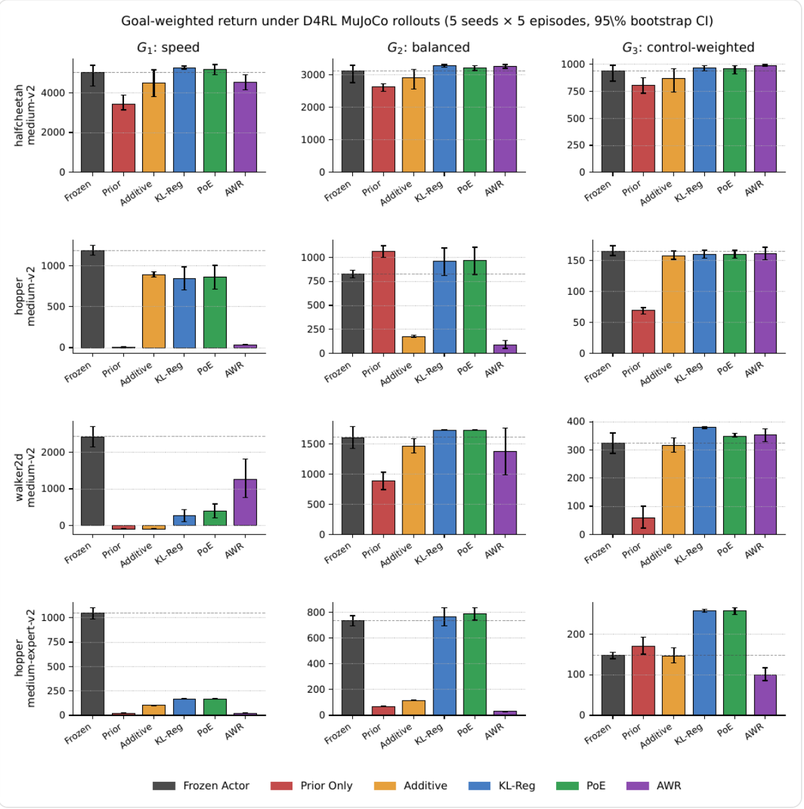
4
Hurt

12 headline cells (4 D4RL settings × 3 goals), ten seeds; five seeds give 4 / 5 / 3

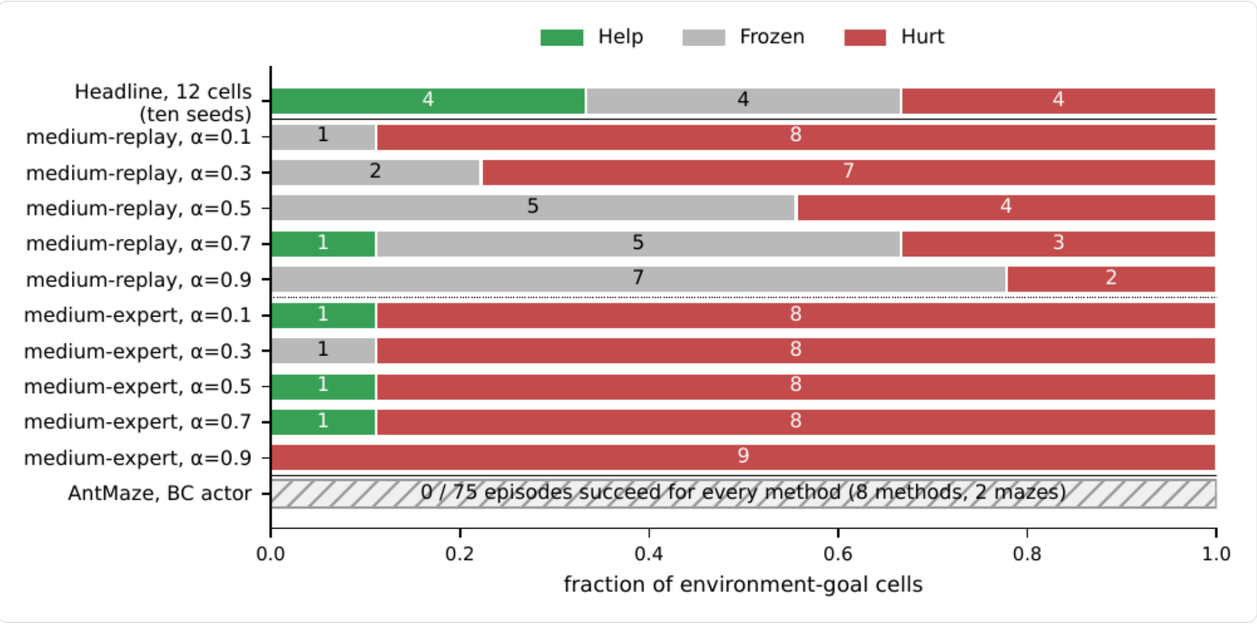
Help: Hopper-M balanced +144, Hopper-ME balanced and control +55 / +110, Walker2d-M control +55

Tie: HalfCheetah-medium on every goal

Hurt: the speed goal, wherever the actor is near its velocity ceiling (−291, −2,030, −878)



Finding 4: actor competence sets the ceiling, not α



Near the ceiling: medium-expert is Hurt in 8 or 9 of 9 cells at every α

A failing actor stays failed: AntMaze BC actors solve 0 of 75 episodes, and so does every method

A competent actor is kept: on Adroit pen, PoE ties the actor while prior-only and additive hurt

Baselines: critics help in places, but keep the boundary

Method (best knob per cell)	Help / Tie / Hurt	Critic	≡ form	Falls back on
PoE / KL-Reg	4 / 5 / 3	No	Yes	The frozen actor, if the prior is weak
CQL-guided	5 / 7 / 0	Yes	No	The critic, errors included
IQL-guided	6 / 6 / 0	Yes	No	The critic, errors included
Critic-gradient step	helps 3 of 12	Yes	No	A raw $\nabla_a Q$ step off the data support

Five seeds, 12 headline cells. With a reliable, validated critic, CQL-guided steering is a strong option. But its best β was picked per cell from realized returns, single-β use is fragile (IQL at β = 1 collapses to about 5 return on hopper medium-expert), and a critic in the action path is heavier to revalidate than one knob.

In practice: deciding before you deploy

A sensitivity score **S**, the closed-form mean KL between the α = 0.5 composition and the actor on dataset states, ranks steering risk with no rollouts. The three five-seed Hurt cells score highest (Spearman −0.77, $p = 0.003$ on 12 cells; −0.67, $p < 10^{-4}$ over 30).

1

Retrain if the actor fails

Near-zero offline success means no reweighting will help.

2

Abstain if S is high

Steering would move the validated actor far from itself.

3

Else steer within a KL budget

Pick the smallest α whose dataset-state KL stays under κ.

The separation is retrospective and imperfect: S ranks risk, but no threshold transfers across regimes.

Takeaway

An actor-anchored safety layer for objective shift, not a universal return maximizer.

PoE ≡ KL-Reg for Gaussian actors, with $\beta = \alpha / (1 - \alpha)$

Stays with the actor when the prior is bad

Bounded by the competence of the frozen actor

paper: arxiv.org/abs/2604.22873
code: github.com/eliashossain001/poe_composition
project: eliashossain001.github.io/poe-steering